Modelo Lineal Generalizado
Guía de R
(Fuente: Faraway, 2006)

1. Tablas de 2×2

Crearemos una tabla en R. Para ello usaremos el comando gl que genera factores siguiendo un patrón determinado.

Así, por ejemplo:

```
>gl(2, 1, 20)
 [1] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
Levels: 1 2
```

```
> gl(2, 2, 20)
 [1] 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
Levels: 1 2
```

```
> gl(2, 8, labels = c("si", "no"))
 [1] si si si si si si si si no no no no no no no no
Levels: si no
```

Para crear la tabla siguiente haremos:

	Edad	
Cantidad	≤ 40	> 40
≤ 20	50	15
> 20	10	25

```
> y <- c(50,15,10,25)
> edad <- gl (2,1,4,labels=c("<=40",">40"))
> cantidad <- gl(2,2,labels=c("<=20",">20"))
> tablaex <- data.frame(y,edad,cantidad)
> tablaex
   y edad cantidad
1 50 <=40      <=20
2 15  >40      <=20
3 10 <=40      >20
4 25  >40       >20
```

Para visualizar la tabla de la forma habitual es útil el comado xtabs:

```
> (ov <- xtabs(y ~ cantidad+edad))
        edad
cantidad <=40 >40
    <=20   50  15
    >20    10  25
```

Para obtener los marginales de filas y columnas podemos hacer:

```
> (mfila <- prop.table( xtabs(y ~ cantidad)))
cantidad <=20  >20
         0.65 0.35
```

```
> (mcolumn <- prop.table( xtabs(y ~ edad)))
edad <=40  >40
      0.6  0.4
```

Calculamos los valores esperados y los estadísticos

```
> (esp<- outer(mfila,mcolumn)*100)
        edad
cantidad <=40 >40
    <=20   39  26
    >20    21  14

> (G<-2*sum(ov*log(ov/esp)))
[1] 22.49692
```

```
> (chi.pearson<-sum( (ov-esp)^2/esp))
[1] 22.16117
```

El test de χ^2 de Pearson también lo podemos implementar a través de

```
summary(ov) Call: xtabs(formula = y ~ cantidad + edad) Number of
cases in table: 100 Number of factors: 2 Test for independence of
all factors:
        Chisq = 22.161, df = 1, p-value = 2.507e-06
```

La corrección por continuidad de Yates que sustrae 0.5 cuando $y_{ij} - \widehat{\mu}ij$ es positivo y suma 0.5 cuando es negativo,está implementada en

```
> prop.test(ov)

        2-sample test for equality of proportions with continuity correction

data:  ov X-squared = 20.1923, df = 1, p-value = 7.003e-06
alternative hypothesis: two.sided 95 percent confidence interval:
 0.2801819 0.6868510
sample estimates:
   prop 1    prop 2
0.7692308 0.2857143
```

Esta corrección da mejores aproximaciones para muestras de tamaño pequeño.

El comando prop.table calculas las proporciones muestrales:

```
> prop.table(ov)
        edad
cantidad <=40  >40
    <=20 0.50 0.15
    >20  0.10 0.25
```

2. Tablas de doble entrada de mayor tamaño

Snee (1974) presenta 592 datos on estudiantes que fueron clasificados de acuerdo a su color de cabello y ojos.Estso datos se hallan en la libreria faraway que puede bajarse del CRAN de R.

```
>library(faraway)
> data(haireye)
> haireye
     y   eye  hair
1    5 green BLACK
2   29 green BROWN
3   14 green   RED
4   16 green BLOND
5   15 hazel BLACK
6   54 hazel BROWN
7   14 hazel   RED
8   10 hazel BLOND
9   20 blue BLACK
10  84 blue BROWN
11  17 blue RED
12  94 blue BLOND
13  68 brown BLACK
14 119 brown BROWN
15  26 brown   RED
16   7 brown BLOND
```

Podemos visualizarlos en forma de tabla mediante el comando xtabs y calcular el estadístico del test de Pearson:

```
> (color <- xtabs(y ~ hair + eye, haireye))
       eye
hair    green hazel blue brown
  BLACK     5    15   20    68
  BROWN    29    54   84   119
  RED      14    14   17    26
  BLOND    16    10   94     7

> summary(color)
Call: xtabs(formula = y ~ hair + eye, data = haireye) Number of
cases in table: 592 Number of factors: 2 Test for independence of
all factors:
        Chisq = 138.29, df = 9, p-value = 2.325e-25
```

Por ejemplo, para realizar un análisis exploratorio podríamos realizar un gráfico representativo y considerar las siguientes opciones cuyos resultados se muestran en las figuras siguientes:

```
dotchart (color)
```

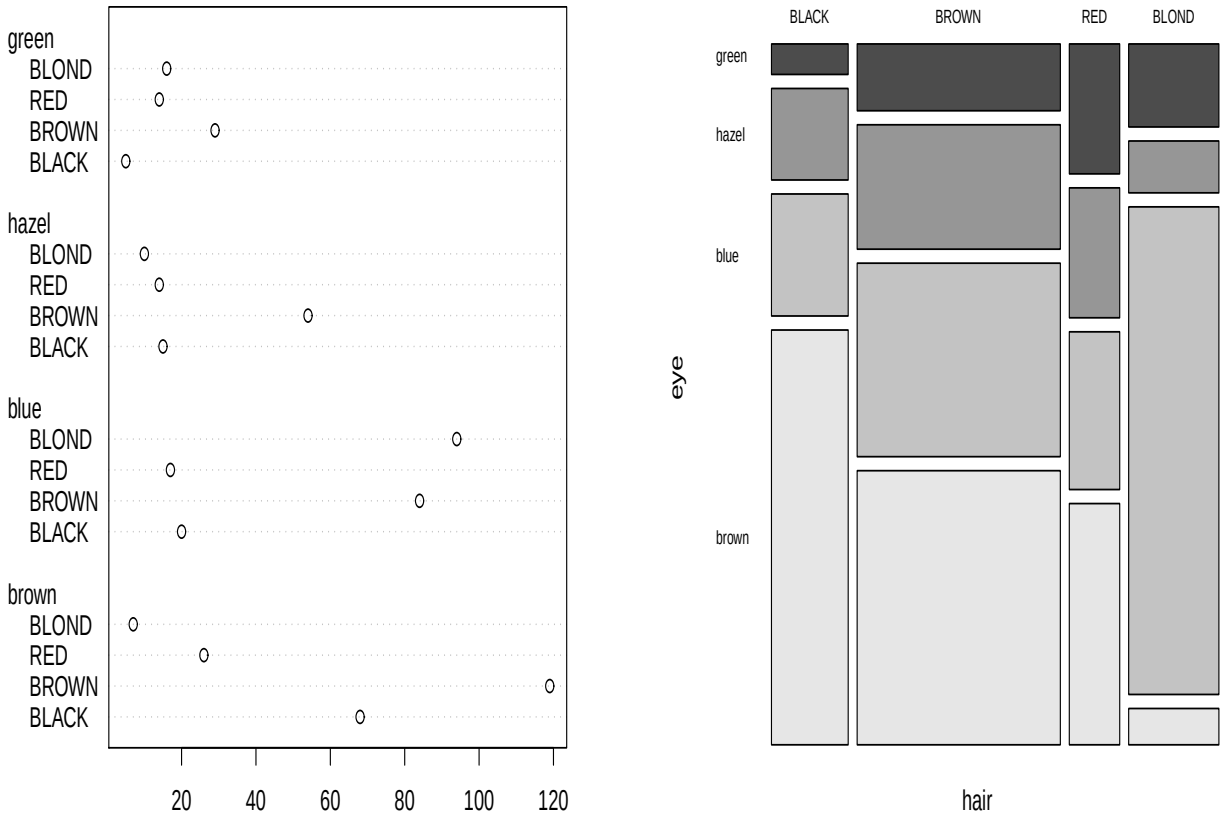

Figura 1: dotchart y mosaicplot

```
mosaicplot(color,color=TRUE,main=NULL,las=1)
```

3. Tablas de 3 variables

Appleton, French y Vanderpump (1996) reportan los datos de un estudio longitudinal realizado durante 20 años a fin de estudiar los efectos de fumar. Durante 1972–74 un estudio mayor que también consideraba otras variables clasificaba a las mujeres de acuerdo a su edad y si fumaban o no. Sólo las mujeres fumadoras son presentadas aquí. Muy pocas fumadoras dejaron el estudio. Aquí presentamos los datos.

```
> (datos <- xtabs(y ~ smoker+dead+age, femsmoke))
, , age = 18-24

      dead
smoker yes  no
   yes   2  53
   no    1  61

, , age = 25-34

      dead
smoker yes  no
   yes   3 121
   no    5 152

, , age = 35-44

      dead
smoker yes  no
   yes  14  95
   no    7 114

, , age = 45-54

      dead
smoker yes  no
   yes  27 103
   no   12  66

, , age = 55-64

      dead
smoker yes  no
   yes  51  64
   no   40  81

, , age = 65-74

      dead
smoker yes  no
   yes  29   7
   no  101  28

, , age = 75+
```

```
      dead
smoker yes  no
   yes  13   0
   no   64   0
```

Si colapsamos por edad, es decir si no tuviéramos en cuenta esta variable, quedaría:

```
> (colapsada.age <- xtabs(y ~ smoker+dead, femsmoke))
      dead
smoker yes  no
   yes 139 443
   no  230 502
```

Si quisiéramos testear independencia en esta última tabla podríamos hacer:

```
> summary(colapsada.age)
Call: xtabs(formula = y ~ smoker + dead, data = femsmoke) Number of
cases in table: 1314 Number of factors: 2 Test for independence of
all factors:
        Chisq = 9.121, df = 1, p-value = 0.002527
```

Vemos que el p-value = 0.002527, con lo cual rechazaríamos la hipótesis de independencia entre estas variables.

¿A qué conclusión llegaría si considera el grupo de edad 45–54 y repite este último test? ¿Parece esto coherente con el resultado anterior?